Rational Choice Theory Considered as Psychology and Moral Philosophy*

PHILIPPE MONGIN
DELTA, Ecole normale supérieure

This article attempts to assess Jon Elster's contribution to rational choice in *Ulysses and the Sirens* and *Sour Grapes*. After reviewing Elster's analysis of functional versus intentional explanations, the essay moves on to the crucial distinction between the *thin* and *broad* theories of rationality. The former elaborates on the traditional economist's preference/feasible set apparatus; the latter is the more demanding theory which inquires into the rationality of beliefs and preferences. Elster's approach to the broad theory normally consists in using the thin theory as a reference point and in making purposefully limited departures from it. The essay illustrates the method while commenting on Elster's discussion of autonomous preferences in *Sour Grapes*. It goes on to stress some important analogies between Elster's use of the thin and broad theories, on one hand, and Weber's ideal-typical method, on the other. The final assessment is phrased in terms of these analogies; it is suggested that Elster is at his best when the ideal-typical method and his own separate from each other, that is, when he comes to grips with the broad theory in its own terms.

I. INTRODUCTION AND PREVIEW

The texts by Jon Elster which will be considered here, *Ulysses and the Sirens* ([1979] 1986) and *Sour Grapes* (1983b) (hereafter cited as *US* and *SG*), are the works of both a philosopher or social scientist well versed in the intellectual methods dear to Anglo-American scholars—

*Received 26 July 1989

A lengthier and slightly different version of this essay appeared as "La théorie du choix rationnel comme morale et psychologie," *L'âge de la science*, vol. 1 (Ethique et philosophie politique), Paris, 1988, pp. 157-97. M. Anspach drafted a translation from which this essay has evolved. The author acknowledges useful comments from I. Creppell, J. Elster, D. Hausman, R. Morrissey, P. van Parijs, and R. Wokler, as well as partial financial support from the ARI "Communication" (CNRS, Paris).

Philosophy of the Social Sciences, Vol. 21 No. 1, March 1991 5-37

logical analysis, argumentation through examples and counterexamples, and mathematical models—and of a French moralist who knows his La Fontaine by heart and can find new ways to compare Pascal and Descartes. In these tightly drawn writings free of all subjective expression, a rare personal disclosure enlightens us as to Elster's singular vocation: He quickly learned, he tells us, that he was neither a novelist nor a mathematician, but his first efforts were not in vain, for "to fail is always to fail at something." Even if one were to accept this judgment at face value, one would have to admire the force of the Hegelian double negation: A truly encyclopedic facility has resulted from these alleged failures. Among his contemporaries, Elster is unmatched in his skill at bringing together the *disjecta membra* of academic culture. For those who do not claim to possess his vast learning, but who strive in their own way to counter the modern dichotomy of the "two cultures," *US* and *SG* ought to be important works as much for what they attempt as what they achieve, which is still open to revision. It goes without saying that an attentive reading of Elster sometimes taxes one's self-esteem. The author of this essay did not at first grasp the magnitude of the task ahead. But he soon adopted as his own the proud opening declaration of *US*: to fail in the face of such delicate analysis is not to fail *simpliciter*.

The two books singled out for special attention among an already substantial and still growing output[1] are collections of articles which have already been rewritten and assembled together. What emerges from these essays is not so much the author's social philosophy (which he has begun to expound in his more recently published *The Cement of Society*, 1989a) as a way of thinking about human affairs in the light of rational choice theories. About this general method there is scarcely anything new. But what makes it attractive in Elster's hands is, for one thing, that the rational choice theories which he employs are somewhat unusual and, for another, that he explores domains which have been neglected by most social scientists who use approximately similar conceptual tools. In the author's experience, this first feature is usually overlooked by Elster's readers. The theories of rational choice to which he is drawn are, roughly speaking, those of the economist but *with a twist*. The twist, for instance his emphasis on the pervasiveness of preference changes, can be easily passed over by the hurried reader. Such a blunder is inescapable if one attempts to force the analyses of *US* and *SG* into a ready-made cleavage between

homo economicus (who is propelled by his own endeavors) and *homo sociologius* (who is pushed by the group); the two books will then be seen as just another tribute to *homo economicus*. Part of the motivation of the present essay was to explore the extent and interest of Elster's contribution to heterodoxy; hence the detailed account of his view of changing preferences.

The second feature, however, is the more important of the two. The *objects* on which the concept of rational choice is supposed to cast light are of an unusual kind, belonging—by and large—to the archaic provinces of the psychology of sentiments and of moral philosophy. *US* and *SG* offer, although inexplicitly, a program designed to reform these fields. Much time will be spent in this essay to make that program explicit and confront it with some standard philosophical perspectives on individualism, intentionality, and rationality. Such a reflective effort was needed, since Elster is avowedly not a methodologist; he is interested much more in examples and mechanisms than in a priori reasoning and broad pronouncements.

The distinction which has proved the most useful to reorganize Elster's contribution is that between the *thin* and the *broad* theories of rational choice, an innovative distinction which is at the core of *SG*. Roughly speaking, the former is the economist's preference—opportunity-set apparatus with various amendments and much flexibility built into it—whereas the latter comes close to a theory of the true and the right. The former relies on a concept of rationality as consistency, whereas the latter explores the meaning of such statements as "it is rational to believe such and such" and "this is an irrational desire." Elster's program in psychology and moral philosophy can only be realized by recourse to a broad theory of rational choice, but his specific approach will be seen to consist in setting *a broad theory which minimally departs from the thin one*. Despite some significant conceptual differences, this attitude is reminiscent of Weber's celebrated discussion of ideal-types. With Weber, too, the scope of the *explicandum* is so wide that it seems hard to assess it adequately without the help of a broad theory of rationality. Whatever may be the meaning of "axiologic neutrality," it seems hardly possible to deal with the textbook example of Weberianism—the Calvinist's faith and his worldly pursuits—without touching on the issue of *justification* of beliefs and desire. There, too, despite a seeming inability to handle such delicate problems, the formal conception of rationality is granted

at least heuristic precedence over the substantial one. As is well known, the "absolute" and "subjective rationality" ideal-types (which come rather close to the economist's characteristic modeling of rational behavior) are to be applied first. If they fail, as they will in the case of the Calvinist, at least that failure will influence our understanding of the so-called residue. Their mode of operation is prototypical of the action of the more substantial, thicker ideal-types which will have to be applied in the second place.

The prima facie analogy of the ideal-typical method (as broadly defined at least, if not in its literal Weberian wording) with Elster's has struck this author as offering a useful perspective on the latter. For one, the ideal-typical method is at present still widely influential among social scientists. When asked about the merits of their models, many economists do not have to be pressed very far to fall back (though perhaps unwittingly) into quasi-Weberian reasoning. They would claim, for instance, that applying what may appear as too strong rational models to behavioral data is at least informationally productive. It is not irrelevant to inquire about the connections between such a well-established method and the dialectic of the broad and thin concepts of rationality. What is done here is after all but another application of the ideal-typical approach itself—the latter is used as a reference point to redescribe Elster's work. For another, and even more important, the dialectic of the two theories of rational choice runs into the same difficulties as the ideal-typical method. The conclusion of this essay is largely devoted to disentangling the former from the latter, and it will be claimed that Elster's results are *least* satisfactory when there is *maximal* entanglement. The objection will be seen to be of an eat-the-cake-and-still-have-it type. Both approaches are attempts to dispense with a full-fledged thick theory while encroaching on its territory. The clever and economical way of making the best of the thin theory before moving to substantial claims on the rationality of beliefs and desires turns out to be productive of *some* results but not as informative as it was hoped. The broad theory must also be addressed in its own terms, as Elster, fortunately if not consistently, manages to do in some of the best parts of *US* and *SG*.

The essay is organized as follows. Sections II and III attempt to capture the philosophy of science underlying Elster's work. Sections IV and V elaborate on the distinction between thin and broad theories of rationality. Section VI discusses adaptive preferences, while Section VII offers a general assessment of the work under review.

II. THE THREE TYPES OF EXPLANATION

In the opening chapters of *Explaining Technical Change* (hereafter cited as *ETC*), Elster (1983a) sets forth his general philosophy of science. At the highest level of generality, he argues for the methodological unity of the sciences by recalling the constraint of the test and the hypothetico-deductive formulation of scientific theories which goes along with it (*ETC*, 15-16). He does not say so precisely, but it emerges from his analysis of intentions and functions that explanation is causal, in a uniform sense of the word, whatever the science considered.[2] At a lesser level of generality, the methodological unity of science is shattered: Disciplines separate from one another not only according to the objects with which they deal *but according to their specific types of explanation*. In physics, explanation is strictly causal, in the obvious sense of mechanical causality. In biology, it is predominantly functional and in the social sciences, predominantly intentional (*ETC*, 17-24). The distinction between these last two modes justifies Elster's recurrent criticism of functionalism in Marxism and anthropology (*US*, 28-35; *ETC*, 35-68).

Since in biology and the social sciences, if not in physics, the nature of the object induces each time a mode of explanation that is predominant but not exclusive, Elster is led (in *US*, chap. 1) to an original discussion of the burden of proof in science: Biology might possibly admit of intentional models as happens in game-theoretical ethology, while economics or sociology might possibly explain functionally (i.e., by the appearance of beneficial effects) how given institutions come to prevail; but in each case, the *burden of proof* must be assumed by specifying mechanisms which ordinarily are not necessary to the judgment that the alleged explanation is well formed. According to Elster, specialists are inclined to content themselves with a *general presumption* of explanation: It is normally enough for a proposed schema to appear logically compatible with some underlying mechanism that is already well known. When one departs from a discipline's preferential mode of explanation, the tentatively adduced mechanism (e.g., strategic interaction of animals, natural selection of institutions) can no longer be left implicit.[3]

The privilege of intentional explanation in social science involves the endorsement of methodological individualism, provided that one adds a premise which Elster does not state explicitly but to which he

obviously subscribes: "Collectives" do not have intentions that are proper to them.[4] Even if it is commonplace in contemporary social science, the unreserved adoption of methodological individualism should be stressed here, coming as it does from an author who, from his earliest work, has persistently interested himself in the Marxist conception of history. In fact, Elster deliberately takes this stand with a program for the renovation of Marxism in mind: "Marxist theory will continue in its stagnant state unless it explicitly espouses methodological individualism" (*SG*, 142). Such a program is pursued in chapter 4 of *SG*—which finally proves aporetic—on "interest or situation-induced beliefs," and above all in Elster's (1985) *Making Sense of Marx*.

The philosophy of science whose schema we have just outlined will doubtless excite few objections. In Elster's view, its most controversial aspect is "the denial of a role for functional explanation in the social sciences" (*ETC*, 20). Without embarking on a long preliminary debate, we must nevertheless point out several difficulties of interpretation. In the first place, Elster combines the usual Popperian or neopositivist thesis of the methodological unity of the sciences with a distinction among three modes of explanation. This combination of two points of view—unity and differentiation—is plainly not in itself problematic, since the first applies to the genus—science—and the second to the species—the particular sciences, distinguished by their object. But it is less economical, philosophically speaking, than the unrefined thesis of the methodological unity of science, since it *appears* to make the analysis of explanation depend on a cosmology, as in the case with Comte and, perhaps, with Mill. Just how far would Elster wish to go in this direction? He doubtless could assert that his tripartition of explanations—causal, functional, and intentional—does not imply, as clearly as may at first seem, a classification of the universe. This initial response will be easier to articulate in light of the following discussion. Elster accepts in the second place Davidson's (1980) conception according to which the reasons for action, that is, desires and motives, are equally its *causes* (*SG*, 2-3; *ETC*, 22). In this context, "cause" must, of course, be understood in a mechanistic and not a finalistic sense: Davidson's thesis serves in particular to replace the supposed causal efficacy of *ends* with the nonteleological eficacy of *intentions*. With Elster, as for Davidson, this thesis comes into play as a last resort. It admittedly has the advantage of ruling out a very direct

form of teleological explanation, but it seems to clash with the radical materialism to which both authors subscribe. It proves indispensable, however, because the reduction of the mental to the physical is impossible and, following another of Davidson's arguments, because it is logically admissible to combine methodological antireductionism with the essential part of materialism.[5] The relation between functional and causal explanation is fortunately easier to examine. For Elster, to explain an organ by its function, in fact by its beneficial effects on reproduction, is to foreshadow a full-fledged causal explanation of a kind which the theory of natural selection can render: "We can use functional explanation in biology because we have a causal theory" (*ETC*, 21). Perhaps the reader has better understood by now how the Elsterian distinction between three modes of explanation must be articulated. While obscurities remain, it turns out that this distinction should not shatter the usual notion of causality and that it may accommodate a form of materialist monism.

In the third place, the critique of functionalism in the social sciences, on Elster's admission the most debatable conclusion to follow from his methodology, is surely beyond dispute if one accepts the simple premise: that to explain is to produce *mechanically* causal links. Since a mechanism linking functions to organs exists in biology, whereas there is no mechanism of comparable generality linking functions to institutions in the social sciences, *functional explanation can be admitted in biology for the same reason that it must be excluded from the social sciences.* This critique is virtually unassailable, and one only regrets a sort of hesitation, an unnecessary tone of caution, in its formulation. Elster sometimes seems less anxious to banish functionalist reasoning from the social sciences than to regulate its use by extending, against Merton himself, the meritorious work of redefinition which the latter had already undertaken in *Social Theory and Social Structure* (1957).[6] In fact, there can be no correct functionalist explanation in sociology, since, if it is correct, the explanation loses ipso facto its functionalist character: It is only complete once the sociologist exhibits a causal retroaction of the beneficial effect on the institution, but then this causal retroaction carries the whole weight of the explanation and one should say simply that *the effect* in question explains the institution; there is no sense in maintaining that *the beneficial character* of the effect plays any explanatory role.[7]

III. INTENTIONALITY AND RATIONALITY

Elster brings the rationality principle into play only in a second phase, as a specification of intentional explanation. Whereas Popper (1967) and his disciples, for instance Watkins (1970), believe that once the concept of rationality has undergone sufficient weakening, it may be applied to the whole spectrum of the social sciences, Elster proceeds in the opposite direction to a position of strategic retreat: "Intentional analysis does not presuppose a rational actor" (*US*, 153). In fact, he *combines* a weakened principle of "imperfect rationality" (*US*, chap. 2), as Watkins terms it, with the thesis of a nonrational intentional residue. The latter is made necessary in his view by the consideration of intrinsically contradictory psychological states: "It is a shallow kind of social science that denies or disregards such phenomena" (*US*, chap. 2). Such states may imply two antagonistic beliefs or desires but also, more strikingly, a *single* contradictory desire, as in the case of *willing what cannot be willed*: One cannot (logically speaking) will to be natural or spontaneous nor will an absence of will nor will to believe (*US* 3.9; *SG* 2.2). One could complete the list of failures of the rationality principle even in its weakened form with the very important use of *failures of human inference*. Elster provides a compelling example of this drawn from a historical study by Laqueur (1980): "While many Germans thought that the Jews were no longer alive, they did not necessarily believe that they were dead" (p. 201). Laqueur's sentence points to an extreme and hideous form of the "framing effect," in Tversky and Kahneman's (1981) sense. A more banal example of failed inference is Bar-Hillel's discovery that individuals tend to overestimate the probability of intersections of events and underestimate that of their unions. Weber's "principle of subjective rationality," with which the Popperians are more or less content, is still too restrictive to permit the social sciences to accommodate these phenomena. Elster is hardly more sympathetic to psychoanalysis than is Popper (cf. "the elimination of the Freudian unconscious as a theoretical entity [is] a highly desirable goal," *SG*, 152), and even if he does not use these words, there is no doubt that he would agree with Popper's (1945) important claim that psychology is a social science like any other.[8] In the present writer's view, such an understanding of psychology readily implies that the social sciences in general can no longer be satisfied with the rationality principle alone.

Nonetheless, this principle retains heuristic precedence. Weber believed that the ideal-type of absolute rationality should take priority in research over that of subjective rationality, itself coming ahead of the ideal-types of nonrational intentional behavior, of nonintentional finalized behavior, and last of causally determined behavior, in that order (pp. 435-36). Elster does not say anything different in *US*. He provides, for example, an interesting illustration of the way in which game theory applies the Weberian precept: The ideal-type of absolute rationality fails when, under the assumption that players have complete information about the payoff matrix, no strategy clearly emerges as the rational solution. Many game theorists would add that this problem arises as soon as the game is not a zero-sum one and none of the players has a dominant strategy. Elster does consider the possibility that users of game theory must run through the entire sequence of Weberian ideal-types, overdetermining their models in the end with a hypothesis of causal determinism (*US*, 156).

He complicates the general theme of successive ideal-types with a variation which is itself traditional: "Economic" hypotheses, that is, those positing selfish rather than altruistic agents, enjoy a certain precedence within the rational approach itself (*US*, 142; *SG*, 10). Put forward with all due precaution, this new rule illustrates *the proliferating aspect of the ideal-typical method*: The choice of hypotheses unfortunately does not consist merely in following Weber's one-dimensional sequence; each step branches out in its turn. Elster brings out very clearly another dilemma which he claims to be coincident with the difference in approach between sociology and economics: "The sociologist typically assumes that there are very large socially determined differences between individuals . . . The economist, on the other hand, would tend to think that preferences are basically similar, and that most of the observed differences in behavior can be explained through differences in the opportunity set" (*US*, 138-39). It may be added that sociology is much less eager than economics consistently to apply the methodological principle according to which temporal changes in preferences should only be postulated as a last resort.[9] We will return to this important point and to the more general difficulty associated with the application of the ideal-typical method.

The interest of *SG* and *US* largely depends on the illustrations and elucidations with which Elster enriches this very common method. Rather than running through the gamut of his analyses and examples,

we would do better to concentrate on the application which seems to be the most original. It has already been said that his two books in effect sketch *a program for the reform of psychology and of moral philosophy*. If the leading idea of that program corresponds roughly to the thesis already quoted from *Open Society*, Elster's challenge is to pursue it. The selected procedure is to study psychological states and form moral judgments mainly (if not exclusively) on the basis of appropriate notions of individual rationality; the ambiguity in this formulation lies ultimately in the word "appropriate," but it already underlies the plural form of "notions." Realizing this program involves one in finding a matrix of rational models that would convey *more* information than the overly general distinction between "absolute rationality" and "subjective rationality" and yet would not end up being saddled with the claims of a specific theory of rationality

Indeed, an attractive solution at this point would be to turn to the economists and decision theorists who, starting from the basic optimization model, develop a taxonomy suited to three typical situations: perfect certainty on the agent's part (this is the most elementary model: the competing firm "takes" the market price); exogenous subjective uncertainty (the farmer does not know with certainty what the weather at harvest time will be); strategic subjective uncertainty (the farmer does not know with certainty how much competitors will supply). As such, this taxonomy is already less summary than the simple opposition between "absolute rationality" and "subjective rationality,"[10] but of course its greatest interest comes *after* the analysis of each component: We all know the remarkable vigor displayed in the wake of von Neumann and Morgenstern's founding work, on one hand, by decision theory (which in fact is essentially preoccupied with exogenous uncertainty) and, on the other, by game theory. Excellent results can be achieved by building on this corpus. But by letting one's research program depend on the preexisting taxonomy of the economists and their consorts, one runs the risk of leaving out, as incapable of assimilation, too great a part of the explicandum. The problem is that the underlying model on which the taxonomy was built, that of optimization in perfect certainty, is, despite appearances, an *extremely determined* model (this crucial point will be made clear later). As a result, the "generalization" contained in the models of uncertainty misses many possible applications; moreover, this "generalization" is often one-sided—it is itself too determined not to need to be generalized in its turn (think for instance of the expected utility model, which

not so long ago was held to yield up the *whole* theory of choice in exogenous uncertainty). Elster is more aware of these difficulties, it seems, in *SG* than in *US*. The latter book offered (in chapter 3) a taxonomy of rational models which is still open to serious objections, whereas *SG*, a more recent book, offers (in chapter 1) a conceptual map which is much more satisfactory for launching Elster's program in psychology and moral philosophy.

This map rests on the threefold distinction between the models used by economists; the "thin theory" of rationality, which will now be distinguished from these models; and diverse variants of the "broad theory." We examine it in the following two sections, before showing how Elster makes it serve the elucidation of the problem of adaptive preferences.

IV. THE THIN (FORMAL) THEORY OF RATIONALITY

The *thin theory* of rationality explains action by reasons—beliefs and desires—the content of which it does not examine. It stipulates only that beliefs are not contradictory and that desires are consistent with one another. "Consistency, in fact, is what rationality in the thin sense is all about: consistency within the system of desires; and consistency between beliefs and desires on the one hand and the action for which they are reasons on the other hand" (*SG*, 1). Microeconomics provides the canonical example of this thin, or formal, conception of rational choice. The model of perfect certainty, to which textbooks give the excessively ambitious name of "consumer theory," includes the following components: an initial set of physically realizable consumption possibilities, and various financial constraints, depending on the consumer, which reduce this set to a "feasible" subset; preferences over the initial set, formalized by a complete, transitive, reflexive binary relationship (i.e., a weak ordering). Roughly speaking, the first set of givens corresponds to consumers' *beliefs*, which are here assumed to be correct, while the second corresponds to their *desires*. The formalism of this model is obvious: The problem of the content of beliefs is not even raised at this stage, and the only constraint that determines the desires is the transitivity axiom, which virtually every writer on the subject (Elster included; cf. *SG*, 6) views as the natural expression of choice consistency. *Action*, in this model, is identified with the basket of consumption items that the agent

chooses, according to the following rule: It is, among the baskets of the feasible set, the best element for the preference relationship. The existence of such an optimum is not self-evident, and the model introduces standard mathematical restrictions to ensure it.

Two features emerge at this point from the microeconomic example. In the first place, it translates the intuitive language of actions and reasons into the precise and more easily mathematized language of *preferential choices under constraints*. In the second place, it offers a mechanism for fitting actions to reasons. By contrast, the connection between the two in the standard definition of rationality is a circular one: Rational action is that "which has reasons"; but the reasons here are not defined otherwise than as the beliefs and desires underlying the action. Doubtless it is specified that the action is *consistent* with reasons, but this expression, like that of the *adequacy* of means to ends in the classical philosophy of action, figures here as a kind of empty space, an appeal to a missing connection.[11] The interest of the foregoing microeconomic theory, and more generally of *optimizing decision theory*, lies in the fact that it fills this space, that it specifies this connection.

The two features just described are separable from each other; in other words, one could conceive of models that resort to the formalism (in both senses of the word) of preferences and feasible sets but that rely on a connection other than that of optimal choice. This is a crucial point. It does not escape Elster's attention in *SG*, where he subdivides the thin theory accordingly: It is possible that, "given the beliefs of the agent, the action in question is *the best* way to realize his desire. Or, more weakly, that the reasons are reasons for the action if it is *a* way of realizing the desire (given the beliefs)" (*SG*, 3). The models developed by Herbert Simon (1983) under the name of "bounded rationality" are an attempt to give life to the formalism of preferences and feasible constraints *with the help of an intentional mechanism other than optimization*, namely, with that of satisficing choice. The Simonian school's objection to the optimizing conception is not easily summarized, as shown by the discussion in Mongin (1986); in any event, Elster did not seem truly to have taken it into account in *US*, since this earlier book, rather inconsistently, referred more than once to Simon's work (pp. 57-58, 62, 135-36) and yet in another passage plainly equated rational action with optimization (p. 113). It is true that this equation is so common in the literature that it may seem innocuous. Some

technical remarks will perhaps be useful at this point to show that it actually rests on a misunderstanding.

What is called "maximization" or "optimization" comes in two versions, one simple and the other complex. We have just recalled the simple one, according to which the optimizing choice refers to the best element in the feasible set relative to the agent's ordering of the choice space. Superficially, the existence of such an optimum appears to be tied to the transitivity of the relationship and therefore to the very hypothesis that choices are consistent. In fact, the transitivity of the relationship is neither necessary nor sufficient for the property that there is a best element in each subset of the choice space. That it is not sufficient is shown by the example of a weak ordering with an infinite number of indifference classes. The properties of feasible sets quite clearly influence the existence of an optimum. Transitivity is not even necessary, as the following simple example shows: x is strictly preferred to y, y is strictly preferred to z, x and z are indifferent.[12] Besides, the existence of a best element in each subset of the choice space depends on the assumption that preferences are complete, an assumption which does not seem to belong to the concept of rationality and which (even more important) has nothing to do with the concept of consistency. According to the complex version, which is also the most common, the optimizing choice is that which maximizes a utility function. This version makes it even more obvious that optimization *exceeds* the notion of rationality. It is not possible in general to represent an ordering by real numbers unless this ordering is *continuous*; and as Elster points out, "Continuity cannot be part of rationality" (*SG*, 9).[13]

Those accepting the foregoing direct argument or the more roundabout one found in *SG*, where various failures of optimization are brought to light,[14] must be careful not to equate "rational choice" with "best choice." Another formula that might seem convincing at first sight turns out to be equally misleading, namely, the one which states not that the agent maximizes preferences but that he or she *strives* to maximize them. This formula could no doubt be defended if desires were conceptualized in a way other than the modern one in terms of preferences, but in that framework it is, quite simply, confused. To choose according to one's preferences does not imply that one refers consciously to one's preferences, nor even that one knows what they are. But striving—to strive to do something relative to one's preferences, for example to change them, or in the present case, to maximize

them—clearly implies knowledge and probably consciousness as well. Elster elucidates this point admirably by distinguishing actions which consist in "doing something" from those which consist in "bringing about something." To choose according to one's preference, to eat a fruit, for example, is typically to do, and not to bring about, something: "I *prefer* the apple. There is no need to go beyond this and add, falsely, that I take the apple *in order to* bring about a certain sensation in my taste organs, or to maximize a certain sensation" (*SG*, 5). On the other hand, there is nothing wrong with the idea that one seeks to maximize one's (monetary) profit, for here the subject consciously envisages the objective, and the action which realizes it is of the "bringing about" type.[15]

What is left of the thin theory of rationality when optimization is not brought into its analysis? It appears as a formal language, a classificatory system, or a "model" in an undefined sense, which specific theories (e.g., microeconomic optimization and Simon's satisficing), as the only truly informative ones, adapt each to its own purposes. This inescapable conclusion needs qualifying, however. In the first place, the terminology of preferences and feasible sets seems to imply by the very way it is structured that *the two are independent*. The preferences should not depend on the feasible set, in the sense, for example, that the preference for x over y, when z is impossible, should not reverse itself when z becomes possible. Revealed preference theory has much elaborated on such an independence, or "exclusion of irrelevant alternatives" condition, and it is clearly in this direction that the most general definitions of rationality-consistency are to be sought. The above property is violated in some of the examples of adaptive preferences which Elster discusses in *SG*.

In the second place, even apart from the foregoing axiom, the model of preferences and feasible sets is not philosophically neutral. It is the vehicle for a theory of action not quite the same as the one assumed by, for example, the more traditional language of ends and means, the modern terminology of problem solving, or even the Davidsonian lexicon of action and reasons, desires, or beliefs. The ends/means representation has too seldom been discussed in relation to the preferences/feasible set model; it is likely that the distinction between doing and bringing about would be relevant here as well.[16] The word "problem" functions as a primitive, unanalyzed term in contemporary everyday and even philosophical language. It is easier to understand what is *already* philosophically determined in the preferences/

feasible set model when a comparison is made with the language of actions and reasons; it is precisely here that *SG* can serve as a test case. It will become evident that in spite of the methodological views expounded in chapter 1, Elster cannot dispense with the Davidsonian lexicon: It is in the latter's terms and no longer in those of the thin theory that he expresses chapter 2—the analysis of "states that are essentially by-products." Elsewhere (Mongin, 1988, sec. 6), we attempted to show that the author was in a sense forced into this choice of language. This was the case because the interpretation of those psychological states which was made possible by the model did not exhaust that of the informal lexicon; something had to get away, a proof of the constraining character of the model. We would be tempted to enlarge this finding into a thesis of broader scope: *One cannot hope to discover an entirely general conception of rationality, since there does not even exist a language in which to formulate it.* Economists have long believed that they had reached in the optimizing models of microeconomics not only a concept but a theory in all due form at the maximal level of generality. This claim must be abandoned, along with the more modest one that consists in assigning a universal scope (for the description of *all* choices) to the preferences/feasible set conceptual pair.

V. THE BROAD THEORIES OF RATIONALITY

The paradigm of the thin theory, microeconomics, is a curious discipline. It arose historically from the "subjective theory of value," and if it has any explanatory power at all, it is because it causally links reasons to individual actions (this point is convincingly argued, e.g., in Rosenberg, 1976). However, it refrains from theorizing the agents' beliefs and desires beyond the formal constraints of consistency. Hutchison ([1938] 1960) pointed out that the microeconomic notion of rationality does not include that of the rationality of representations or of expectations (pp. 86-88); the recently constituted theory of "rational expectations" makes it now necessary to qualify this statement, but it remains arguably correct.[17] It would appear, then, that microeconomics is on the wrong track: How can one claim that reasons are the causally effective factor while refusing to examine them? Hutchison concluded that the paradox is hopeless: The discipline would seem doomed merely to string together "tautologies." Indeed, the thesis

that it has nothing determinate to contribute recurs constantly in the literature in one form or another. The truth is quite different. The empirical and possibly even testable content of a microeconomic model of perfect certainty derives, broadly speaking, from three sources: (1) the preferences/feasible set model, possibly overdetermined by axioms of revealed preference theory such as "exclusion of irrelevant alternatives"; (2) the twofold condition of consistency with regard to beliefs and to preferences, which is a formal condition though not an empty one; and (3) the optimizing connection between actions and reasons. On this foundation, it was possible to build a theory. That is what economists have done since the end of the last century, without always knowing very clearly whether they were developing necessarily true and informative theorems (either in the Leibnizian sense of necessary truths or in the Kantian sense of the synthetic a priori), necessarily true but empirically empty theorems (that is, the neopositivist stand taken by Hutchison), or empirical propositions (by and large the point of view defended here). In any case, they agreed on the idea that their system had to be built *deductively*, and that the chosen basis, whatever it was really, discharged them from having to refer to psychology or to any other preexisting discipline. This method had immense advantages and quite visible drawbacks which cannot be examined in detail here. But it is clear that it is only one way, and in a sense the most paradoxical one, of founding a social science on rationality.

What Elster calls a *broad theory* of rationality is one which claims to include rationality of beliefs and desires. A broad theory will be much more markedly normative than the preceding one. It will rest on the notion of right judgment, as far as beliefs are concerned, and of autonomy, the Elsterian formula for *good* desires (*SG*, 16, 20). The difficulty with this approach is obvious: How can it be followed without going into the general theory of the true and the good? Elster proposes a middle term:

> Between the thin theory of the rational and the full theory of the true and the good there is room and need for a broad theory of the rational. To say that truth is necessary for rational beliefs clearly is to require too much; to say that consistency is sufficient is to demand too little. Similarly, although more controversially, for rational desires: the requirement of consistency is too weak, that of ethical goodness too strong. (*SG*, 15)

This middle term is the essence of Elster's program: It will mean taking on the areas which microeconomics—and by the same token, any variant of the thin theory—are incapable of handling, for want of a sufficiently large conceptual base, but this will be done while departing as little as possible from this base through deviations from the elementary schemas rather than through radical innovations. This is the way in which Elster intends to introduce traditional psychology and moral philosophy into social science.

This program implies using as far as possible the language of preferences and feasible sets, which involves a curious and fruitful retranslation of classical authors. For instance, the Cartesian maxim, "ne suivre pas moins constanment les opinions les plus douteuses, lorsque je m'y serais une fois déterminé, que si elles eussent été très assurées" (1963, *Discours* 3.594-95), gives rise to three successive interpretations among which Elster attempts to arbitrate (*US* 2.4). The use of a clear and unified language makes it possible not only to *reread* Descartes, Pascal, Stendhal, La Fontaine, or Emily Dickinson but elaborate *taxonomies of real situations*, for example, to distinguish among several forms of precommitment (*US*, chap. 2) or to clarify the relationship between adaptive preferences and other voluntary or involuntary forms of variations in preferences and feasible sets (*SG* 3.2). Beyond these initial results, the aim would be to establish *laws*, if only approximate ones. However, the research has not yet reached this goal, except in particular cases, and even then most often in negative or existential form. It remains to be seen whether this state of affairs corresponds to a particular stage in the undertaking or, on the contrary, to a structural weakness. Finally and most important, the linguistic and taxonomic clarification prepares the way for *ethical judgment*. We said that Elster was a moralist at heart. For example, he wants to apply his distinction between adaptive and counteradaptive preferences to social choices (*SG* 3.4). Thus in the area of individual morality, his probing discussion of states that are essentially by-products brings out more clearly the boundary between the possibility and the impossibility in principle of manipulating oneself (*SG*, chap. 2; *US*, pass.). The Elsterian program obviously requires moving *from the formal to the genetic*: This implies investigating the strength of the arguments and data on which beliefs rely, a well as the rational (that is to say autonomous) or irrational ways in which desires are formed. As far as beliefs are concerned, the task is so broad that it

covers the entire project of an epistemology, and by Elster's own admission, it remains unfulfilled (*SG*, 16-17). Before proceeding to the other more fully worked-out part of the program, it may be interesting to indicate a middle road that has yielded some results. It is not inconceivable to construct a *formal* theory not only of the rationality of *choices* but of the rationality of *beliefs*. Currently, game theorists are advancing in just this direction, which had been advocated at an early stage by Harsanyi (1977). Roughly speaking, a player's representation of one's opponents is taken to be determined by such elementary constraints as A cannot attribute to B a behavior that A would not display in a comparable situation, A knows that B cannot attribute to A a behavior that B would not display, and so on. These constraints can be likened to the symmetry principles of physics; it is also the case that they embody a natural extension of the rationality—consistency concept to the case of strategic uncertainty. They often make it possible to derive surprisingly precise results such as the "no trade" theorems derived from the "common knowledge" assumption.

VI. ADAPTIVE PREFERENCES

The analysis of *adaptive preferences* and of *autonomy* is perhaps the most interesting application of the broad theory in Elster's writings. The thin theory, whether optimizing or not, can only arrive at determinate outcomes by assuming relative stability of one of the two components of the explanation: preferences or feasible set. In effect, the theory's principal results consist in functionally linking—as in microeconomic demand theory—changes in the chosen action to presumed changes in feasible sets. Quite obviously, the linkage is a functional one only if preferences themselves do not change. Having said this, we should note that the thin theory of rationality can accommodate *some* temporal variability in preferences: The theory can always be applied *successively* to the configurations of choices presumed to be stable. But the phenomenon of variability taken in itself can only be analyzed from within the broad theory. The latter then brings into play a multidimensional taxonomy (*SG* 3.1, 2) which we will now restate.

Preference changes can be adaptive, by bringing the desirable closer to the possible ("Ils sont trop verts et bons pour des goujats"), or counteradaptive, by pushing them further apart ("It's always

greener on the other side of the fence"). They can be irreversible or reversible; two particularly interesting cases of irreversibility are habit and learning, which outwardly resemble one another. For example, is the farmer who once worked in the city and who returns nostalgically to the countryside an addict or an informed individual? Elster also makes the following point, which will prove to be important later on: Preference changes that are "causally induced," as in the case of La Fontaine's fox, are *not only ethically but psychologically* different in nature from "intentionally engineered" preferences (*SG*, 110, 117-19). This latter case, schematically represented by the Stoic or Buddhist philosophies, can be analyzed, unlike the former, as a second-order preferential choice, a feature which Elster claims should not be without observable consequences (*SG*, 113-19).[18]

In addition, all the aforementioned cases should be separated from that of *precommitment*, as when Ulysses asks to be tied to the ship's mast in order to resist the allure of the sirens. Here, however, the distinction is external rather than internal to the phenomenon of preference change. In this case, the agent deliberately modifies one's own *feasible set* in such a way that, *one's preferences being given*, some decision becomes unavoidable or, on the contrary, is avoided. Simple as it is in its principle, the external distinction is often difficult to apply: The phenomena of precommitment and of adaptive preferences interfere every time that deliberate modification of the feasible set triggers an adaptation of preferences (through a mechanism which then is not intentional). Elster subtly illustrates this complication with the example of marriage (*SG*, 114-15). In another example, the soldier who asks to be sent to the front acts like La Fontaine's fox and like Ulysses *at the same time*. This complication reflects back on the distinction between determined or intentional changes in preferences: Does the soldier wish merely to place himself in a combat situation, or does that soldier also hope to enhance his own courage in the fire of battle? In this last case, the mechanism of adaptive preferences would be *put to the service* of character planning.

This classification (which can only be rendered schematically here) is illuminating, but certain difficulties remain unresolved. In the first place, Elster's use of the thin theory terminology is not completely rigorous. He should distinguish from the outset between the dependence of preferences on *states* and the dependence on *feasible sets*, instead of calling on this distinction belatedly to clarify matters (*SG*, 122). This would allow him to break the false symmetry between the

typical example of adaptive preferences (La Fontaine's fox) and his alleged example of counteradaptive preferences: the man who, when in London, wants to leave for Paris, and once in Paris, wants to leave for London (*SG*, 111). The fable of the sour grapes can be understood in more than one way, but the most plausible involves a dependence on the feasible set (the fox cannot reach the grapes) rather than on a particular state of the world (the grapes are at a certain height); Elster agrees (*SG*, 122). On the other hand, the Paris-London example requires the opposite interpretation: Counteradaptation can only relate to the state (the city in which the agent is located), since the feasible set remains the same in either state (the agent can always leave one city for the other). The true counterpart to the fable of the sour grapes is found in another saying that Elster quotes: "It's always greener on the other side of the fence." The pair "state/possibility-dependency" ought to have been crossed with the pair "adaptive/counteradaptive." The confusion in taxonomy seems to be rooted in the peculiar features of the chosen examples: In the case of Paris and London, a state (to be in Paris) seems—incorrectly—identical with a shape of the feasible set (to be able to stay only in Paris). More abstractly, it would be important to consider, in the manner of ordinary decision theory, sets E of states of the world, A of possible actions, and C of consequences; the preference relation is defined on C and depends either on A or on a state e realized in E—or, more generally, on both factors.[19]

In the second place, all the cases of possibility-dependency, as well as the definition of autonomy that Elster sketches later on (*SG*, 130-31), raise the following problem: The agent is supposed to evaluate—either in utility or preference terms—some "consequences" which he or she knows are not realizable. To use the terminology of the preceding paragraph, there are no technical difficulties, in supposing that the changing evaluation relates to the whole of C even though A and the given e in E are normally associated with a proper subset of C only. Such an assumption would generalize the traditional example of the consumer, where a distinction is made between the initial set of (physical) possibilities and the various (financially) feasible subsets, while the preference relation refers to the *whole* of the former. It is certainly reasonable to suppose that consumers can evaluate baskets of goods that are beyond their reach. But is the example susceptible of being generalized? Elster discusses the effects of the industrial revolution on preferences *SG*, 133-34). For the purposes of his demonstration, he assumes that the variables which are relevant to the two

utility functions (i.e., utility before and after the Industrial Revolution) are the preindustrial situation x, the industrial situation effectively obtained y, and an industrial society z more equitable than y. He also assumes that the postindustrial evaluation of y depends on that of z: taking z into account downgrades y. Now, the postindustrial evaluation of z is not necessarily available, since it relates to a situation that is logically possible but not feasible.[20] Is it reasonable to assume, as Elster does, that the agents can evaluate, and even perhaps evaluate within a cardinal preference map, what is simply a logical possibility? In the case at hand, Elster may be right to answer in the affirmative, but one feels uneasy with the lack of guidelines in the general method. Despite its evident weaknesses, standard utilitarianism, which would be satisfied here with comparing ex ante utility (preindustrial evaluation of the preindustrial situation) with ex post utility (the postindustrial evaluation of the postindustrial situation), has the advantage of being *informationally much more economical*. It does not make the social choice depend on data that may be impossible to pin down.

The problem encountered here has two sides: for one, the evaluation of simple logical possibilities; for another, the stability of the consequences set C which the agent has to evaluate at successive periods. In terms of the latter, the difficulty is symmetrical to the one which confronted standard applications of the thin theory, in which stability of preferences had to be assumed. Elster can dispense with this premise at the price of requiring that *successive evaluations* relate to one and the same collection of facts; that is, he requires that the pre- and postindustrial utility functions be defined on the same domain—a puzzling assumption in this context, since it endows the preindustrial agent with a premonition of the society to come. A similar problem clearly underlies his example of the concentration camp, illustrating one of his formal criteria of autonomy (*SG*, 131).

In the third place, this last concept is not defined rigorously. The formal criteria that are intended to capture autonomy cannot necessarily be met *simultaneously*. They do appear to conflict with each other in the—pivotal—example of the Industrial Revolution. Elster suggests that the preference change which accompanied this period was a step forward as regards autonomy (*SG*, 134): It does indeed conform to criterion 1 which makes for a presumption of autonomy when the feasible optimum (y) is not the absolute optimum (z), but on the other hand, it violates the criterion put forward (tentatively, it

is true) on page 131 of *SG*, which rules out a reversal of the strict preferences x and y.

At the broader level, the discussion of adaptive preferences exemplify difficulties which doubtless are typical of Elster's *middle road*. The point of the method is to use the conceptual apparatus of the thin theory as far as possible. But when the insufficiency of this theory becomes apparent, it is not always clear what Elster will introduce to supplement it. The distinction between preference changes through habit and through learning leads to using a "divided self" concept (p. 121), which is obviously relevant but also sketchy and, what is more important, antagonistic to the minimalist notion of rationality-consistency. More subtly, when he writes, "Whereas adaptive preferences typically take the form of downgrading the inaccessible options, deliberate character planning would tend to upgrade the accessible ones" (p. 119), Elster once again goes beyond the limits of the thin theory without indicating clearly what should replace it. As accurate as the observation may appear to be, it is a difficult one to translate into the language of ordinary utility theory, since the latter does not include an absolute zero point.

VII. ASSESSING THE PROGRAM

Having described Elster's approach as a variation on the ideal-typical method based on a subtle development of the intentionality principle, we examined in particular its application to individual psychology and moral philosophy. In this area, one may expect novel results from a heuristic which has otherwise been extensively tested on the analysis of collective phenomena. Is it possible now to tally up the successes and failures of the Elsterian program in moral reasoning and psychology? The conclusions of *SG* are probably subject to revision, as were those of *US*.[21] They may occasionally seem lacking in bite. As the reader has by now understood, Elster is a lover of precision, and social science in his conception is more closely akin to an engineering enterprise than to the renewal of a metaphysical project. The pointillism of this approach is intrinsic to it: Faithful to a thesis which has been forcefully argued by Popper or Hempel, it accepts the fundamental banality of our nomological knowledge of humankind and expects the unexpected to come only from confronting particulars. With these two reservations in mind, it should be possible, if not

to draw up a balance sheet of successes and failures, at least to take note of the most typical and most promising results.

We have stressed that his method casts a new light on past writers. Historians of philosophy, with all their justified respect for "internal problematic," will profit from reading Elster's analyses of Descartes, Leibniz, or Marx. To bring out the movement of a text, the artfulness or ambiguity inherent in a given formulation—we quoted as an example the Cartesian maxim from chapter 4 of the *Discours* (1963)—the wisest policy is not always to reapply the author's conceptualization; the choice of conceptual reference points that are *external* may be the best way of bringing out an "internal problematic." Such is the case when the reference points are structured and impose, at least locally, a stronger coherence than that found in the text. They then operate exactly as the ideal-type of absolute rationality on insufficiently rational behavior: The perceived distance between the material and the model (which thus signifies at the same time "norm" and "structure") will yield useful information, and this information will *then* feed into an assignment of meaning that could be called subjective or "internal." The goal is therefore the same as in traditional history of philosophy, but it is attained obliquely, and not by pure reflexivity, in a process which appears not only fruitful in view of the result but commendable in that—as Weber (1922) rightfully insisted—it allows for a form of objective control.[22]

Elster further excels in the invention of *taxonomies*, whether abstract (as in the map of rational models) or concrete (as in the typology discussed in section VI). Here again, we need to emphasize that his approach conforms to significant precedents. Psychology and moral philosophy have traditionally meditated *cases*. If this way of proceeding has sometimes fallen into discredit, that is because it was not always adequately grounded in rational taxonomies (although the rambling style of old-fashioned casuistry has doubtless been much exaggerated to make it a more fitting target for sophisticated wit). If the majority of epistemologists are to be believed, one of science's humbler functions, classification, is nevertheless called on to play a key role in the *Geisteswissenschaften*: It is a natural method to apply by disciplines which have just been claimed to become truly informative only in relation to particulars. One could also attempt to rehabilitate taxonomies without immediately referring to social science. Those which are truly satisfying implement, at least implicitly, a previous nomological knowledge, of which they then provide illustrations,

following the explicans-explicandum schema of traditional philosophy of science. What is scientifically interesting in the classification of animals is the use it makes of, for example, causal relations between organs and functions, between the presence of scales and thermal regulation, and so forth. This way of salvaging the taxonomic approach is well known (e.g., Hempel 1965). It should lead to a distinction between *taxonomies of pure application*, which bring into play well-established nomological knowledge, and *test taxonomies*, which are on the contrary a way of putting underlying laws to the test. Such a distinction is easier to make in the abstract than in practice, where it will doubtless appear a matter of degree. But supposing that we accept it in spite of its probable flaws, and coming back now to social science, we should be able to elaborate on the case made earlier for the use of classifications: Social science must seek *taxonomies which relate to intermediate-level generalities*. These are the only ones which could resemble test taxonomies in the Geisteswissenschaften, granting the fact that ultimate laws are out of the reach of testing. What can be said in this respect about the distinctions and maps which emerge from *US* and *SG*?

Elster is not satisfied with the all too general taxonomies to which the philosophical literature on rationality has too often confined itself: the distinctions between absolute and subjective rationality, between parametric and strategic forms of uncertainty, and so forth. More than once, Elster reaches the stage of test taxonomies, or more exactly, the threshold of this stage. We said that *SG* (3.2) connected an a priori distinction concerning the mode of variations in preferences and a difference, which is independently identifiable and perhaps even observable, between two psychological phenomena: Causally induced preferences would tend to "overshoot," while intentionally combined ones would not. Here, the intermediate level of generality appears in an explicit form. Unfortunately, the test is missing. It would be complicated by the already mentioned problem, that Elster passes over quickly, of frequent interaction between the two modes of preference changes. The soldier who gets himself transferred to the front and becomes excessively impervious to suffering and danger cannot be cited as evidence for or against the conjecture, since this soldier represents a theoretically impure case. Elsewhere, *SG* goes some way toward test taxonomies without reaching them: Neither the underlying generalities nor even the classificatory principles appear clearly. We are thinking of Elster's discussion of the feasibility of intervention in by-

product states. He is divided here between a general impossibility thesis and objections suggested to him by the particular structure of the causal chain leading to the desired state.[23] The question is finally left in abeyance, and one may wish that the analysis had not proceeded by simply juxtaposing conflicting examples or arguments: It might have been possible to classify the by-product states more rigorously and then to examine how the taxonomy was related to the problem of intervention.

To give one last example, *US* sets forth the thesis that human intentionality results in *global* optimizing choices, whereas biological functionality only has to do with *local* optimization—it simulates intentionality in an impoverished fashion (*US*, 9-11). This statement is important in several respects: On one hand, it has a methodological side to it, since it serves to specify relations between the social and biological sciences, as well as the possibilities for legitimate borrowing from one to the other; on the other hand, it can be taken as an empirical generalization, cast at a high level no doubt, but not totally devoid of informative content. The idea would doubtless have to be made considerably more precise: Elster ought to indicate the *maximanda* variables of the two entities that he compares, the human subject and the theory of natural selection, rather than opposing "global optimization" to "local optimization" in the abstract, which has no clear meaning.[24] With this accomplished, one could envisage a form of test (over very general *classes* of events, that is). But Elster does not set off in this direction, and in *SG* he even seems to neglect this earlier thesis, central as it was to *US*. The reason for this may be that the later book draws a much sharper distinction between rational choice and intentionality, optimization and rational choice. The example just taken up is therefore one of a nearly entirely unspecified taxonomy. Once optimization is taken away, nothing remains but the distinction between "global" and "local," which by itself is not enough to elucidate the difference between biological functionality and human choice.

Supposing that we have succeeded in accurately summing up these typical results of the Elsterian method, the empirically oriented reader will no doubt be left somewhat disappointed. Two points still need to be clarified which should lessen the disappointment. First of all, Elster is not only a psychologist but a moralist, and there are times when it is impossible to examine both aspects of his method simultaneously. Elster is a long way from Kantianism, in the sense that he bases his

ethical appraisal not on a priori principles alone but on the nomological structuring of reality. His morality is at the same time an anthropology, and such a presupposition, added to the natural ambiguity of the ideal-typical method, explains why we have seldom felt it necessary in this article to distinguish between the psychological and normative sides of the program. We have in fact constantly emphasized the former since it appeared to be a preliminary condition to the implementation of the latter. This way of looking at things must not cause us to lose sight of the context. The reader might find the taxonomy of preference changes disappointing because it does not clearly lead to empirical scientific results; it is, however, roughly sufficient for elaborating the concept of autonomy on which the moral discussion turns.[25] If we become less exacting in this connection, that is because the prescriptions of law and ethics are to a large extent limited to relating *intensional* notions to one another. And from this point of view, the distinction between two main types of adaptive preferences can, at the level of specificity of *SG*, offer a satisfactory explicans for autonomy. Naturally, such an interpretation of normative discourse passes the bulk of the difficulties on to casuistry or jurisprudence, which will have to confront the notions' extensional aspect. In the second place—this point will doubtless strike one as more decisive—Elster's results are largely determined by the weaknesses of the ideal-typical method itself. In spite of everything that recommends it to common sense—it is in a way *the* method of common sense, as Popper says in substance—it presents formidable drawbacks which the philosophical literature has not adequately brought out. *The ideal-typical method does not easily lead to the formulation of intermediate-level laws.* It remains largely a procedure for interpretation and classification in the already defined sense of application taxonomies rather than test taxonomies. While we cannot attempt to justify these critical claims in detail here, we can present a few arguments.

The method's general idea consisted in making information appear *through difference*—in the case at hand, through the discrepancy between decisions effectively taken and rational models purposely chosen to be too demanding.[26] The aim was to bring out the content of a *positive theory of choices*, either nonrational or rational in a weak sense. The putative regularities were supposed to present themselves more or less spontaneously as soon as the general characteristics of the choice situation were related intelligently to the decision procedure most often associated with them in practice. As it turned out, this

programmatic schema functioned poorly because the rational models placed at the starting point did not provide adequate constraints. More precisely, these models proved to be both *normatively demanding*, which was expected by the proponents of the method, and *empirically loose*, which doubtless was not.[27] The algorithm lost its iterative character, converging every time on one of the first ideal-types put to the test. Returning to Max Weber's hierarchy of types of rationality, it is clear that "subjective rationality" should suffice for nearly all the historical or ethnological explanations. If one now adds, as Elster does, more precise distinctions within the types themselves, and if, for example, one decrees at every successive level that optimization and stability of preferences have lexicographic priority, one will observe once again that the algorithm immediately converges to the "subjective rationality" type, in the lexicographically privileged variant. It will perhaps be said that this variant is especially hard to refute, and that another lexicographical choice might better demonstrate the method's potential. Such is not the case: In the variant that may be called, roughly speaking, "sociological" rather than "economic"—where exogenous determination of preferences is admitted—it is not essentially more difficult to account for the explicandum. Of the many attempts made here and there to elaborate the method,[28] not one, as far as we know, has succeeded in overcoming its general drawback: *the difficulty in testing the model taken initially as a reference.*

These schematic considerations make it clear why the ideal-typical method is in the end more useful for interpretation (understood as attribution of subjective meaning) and classification (understood as taxonomies of application) than for the discovery of laws. The first model considered will usually be the "right" one, and applying the method then comes down to applying this particular model. The laws that it sets in motion more or less explicitly will suffice to derive the explicandum; there is no incentive to seek out a *new* theory. The paradox of the ideal-typical method is that it could only reach its global (heuristic) aim while failing locally (as a source of explicans): Putative regularities were allegedly found *residually*, that is, after examining those parts of the historical or sociological material which had resisted the models. With a few reservations,[29] this was an a priori admissible paradox. The problem is that the paradox was not even put into practice: The method succeeded locally; its models proved to be the "right" explicans. Even worse, the *local* success was taken to be a *global* success. There are many social scientists today who content

themselves with applying models, believing all the while that they are following the ideal-typical method.

One aspect of this process is, of course, well known: Models of rationality have invaded the entire field of the social sciences, pushing out toward the edges the weaker notion of intentionality. This result would have surprised Weber, for whom, manifestly, the right explanations are quite often to be sought beyond the rational type—in his *Protestant Ethic*, he demonstrated this with an obscure and fascinating example. One must be grateful to Elster for having defended the role of an intentionality principle in psychology and moral reasoning. Since he does this while maintaining the heuristic privilege of rationality, he would seem to be very close to the Weberian source—an exponent of an ideal-typical method restored to its original complexity.

In fact, Elster is not exclusively Weberian—far from it. He is able to assign an *effective* role to intention only because he does not always apply the ideal-typical method. For instance, in his analysis of by-product states he does not put rational models to work on behavior. Less obviously perhaps, the analysis of adaptive preferences also violates the canons of the ideal-typical method: It doubtless uses as far as possible the terminology of the thin theory—preferences and feasible sets—and an accompanying notion of rationality-consistency, but Elster is careful enough not to tie himself to any *specific* model of the thin theory, since he does not impose on himself the constraint of preferences stability. This reminder suggests that we should not be content with equating Elster's approach with a mere variation—even a subtle one, even a more genuinely Weberian one than most—of the ideal-typical method.

The "map" of rational models, which was detailed in sections IV and V, ends up playing a double role. On one hand, it supplies the analysis with models, in the strong sense: The thin theory will include a precise axiomatic system, including a stability of preferences assumption; and the broad theory will consist in weakening these various axioms while preserving the basic terminology. In this interpretation of the map, one goes *from the most determinate to the least determinate*. It follows a typically Weberian orientation. On the other hand, it moves from elementary and, in principle, universally acceptable postulates of rationality—those of consistency—to substantial, more normatively loaded and more debatable postulates about the rationality of beliefs and desires. In this interpretation, the map is

oriented *from the least determinate to the most determinate.* Any remnant of the ideal-typical method disappears here. To reason within the framework of the broad or thin theory no longer means confronting a given model with reality but only advancing psychological or moral theses with varying degrees of normative commitment. Elster's program thus hesitates between two possibilities: the indirect approach established by Weber, and another one, more difficult to describe a priori, which theorizes objects directly in the light of successive philosophical postulates. Elster's second approach is not the less interesting. But what else is it—elaborated case by case and felicitously expressed in modern language—than the very theory of the right and the good which *SG* wanted to do without?

NOTES

1. Aside from *Explaining Technical Change* (1983a) and *Making Sense of Marx* (1985), which will be mentioned on occasion, Elster wrote *Leibniz et la formation de l'esprit capitaliste* (1975), *Logic and Society* (1978), as well as two textbooks and numerous articles in English, French, and Norwegian. He is also responsible for several collections of essays. Since this article was written, Elster has pursued the issues of rational choice theory in *Solomonic Judgments* (1989b) and published *The Cement of Society* (1989a).

2. Causality in Elster appears as the explicans of explanation in contrast to the approach of the deductive—nomological model of explanation (where the converse is usually taken for granted). Elster does not take sides on this classic construct.

3. This analysis may complement Hempel's (1965) idea of "explanation sketches" (which was admittedly put forward in a different context). Made suitably precise, it could also serve for a nonsociological interpretation of Kuhn's cliché about "normal science." For a precise discussion of the examples given here, cf. *US* 1.4, 5 and *ETC*, 55.

4. Agassi ([1960] 1973) has shown the crucial role of this premise for the distinction between holism and methodological individualism. In classic fashion, Elster claims that the latter approach is endorsed by the metaprinciple of *reductionism* (the tendency "to seek an explanation at a lower level than the explicandum," *ETC*, 23).

5. Here, materialism refers to the view that every mental or psychological state is in fact a neurophysiological state. Reductionism would refer to the explanation of mental predicates by neurophysiological predicates. The alleged compatibility between materialism and antireductionism in Davidson (1980) seems to the present writer crucially to rest on the distinction between (atomic) *states* and (relational) *predicates* (or *properties*).

6. Merton rejects the implied view of functionalist anthropology: *Every* social phenomenon has beneficial consequences which can explain it. The principle to which Merton himself adheres would be as follows: "*Whenever* social phenomena have consequences that are beneficial, unintended and unrecognized, they can also be explained

by these consequences" (*ETC*, 57). Elster criticizes this principle because it dispenses the social scientist from specifying the causal retroaction of the effect on the explicandum phenomenon.

7. In other words, the characterization which Elster provides (*US*, 29; *ETC*, 57) of a correct functionalist model in the social sciences seems to be *redundant*. Once it is postulated, the feature that Elster denotes by (5)—the institution maintains itself through a positive feedback of the effect it produces—suffices entirely for the explanation. Merton's intuition was that (5) could be deduced, in one way or another, from the other features in Elster's list; if this intuition collapses, it becomes superfluous to examine them.

8. Cf. Popper's (1945) statement: "Psychology—the psychology of the individual—is one of the social sciences, even though it is not the basis of all social science" (2:97); see also Popper (1972, 22nd thesis).

9. Becker (1976) elevated stability of preferences to the rank of a postulate of the "economic approach to human behavior" (p. 5).

10. The above-mentioned articles by Popper and Watkins hardly go beyond this distinction. Weber made something of the further distinction between strategic and exogenous ("parametric" in Elster's word) uncertainty. Defining strategic contexts is not an easy task. The following demarcation criterion may be submitted: Uncertainty is strategic rather than exogenous when it is rational *for me* to take into account *the opponent's* expectations regarding *my* behavior (and not just when it is rational for me to form expectations on the opponent's behavior, nor, of course, when it is rational for me to take into account his expectations in general). If there is something to this criterion, Stackelberg's model of duopoly does not belong to the realm of strategic interaction. Note also that the typology of uncertainties appears to be sensitive to the chosen concept of rationality—a connection which has not received the attention that it deserves.

11. The same could be said of the Weberian notion of *zweckrational*, which is insufficiently explanatory and notoriously difficult to distinguish from *wertrational* (cf. Aron [1936] 1969, 306).

12. The example is borrowed from Sen (1970, 3). Although this is not a transitive relationship, each nonempty subset of $\{x, y, z\}$ admits of a best element.

13. The empirical interpretation of continuity hypotheses is notoriously tricky; in the case at hand, it relates back to a property of the feasible sets as much as to the texture of preferences, which suffices to show that it goes beyond the notion of rationality.

14. One should perhaps distinguish more clearly than Elster does (*US*, chap. 2; *SG*, chap. 1) between these different causes of failure: (1) the concept of optimal solution is clearly defined, but (a) there does not exist an optimum, or else (b) there exist several optimums; or (2) the concept of optimal solution is not clearly defined, either because (c) no satisfying concept can be found, or, more likely (d) several can be found, competing and up to a certain point incommensurable with each other. The problem is notably to untangle (b) and (d), which is not always easy in game theory, where the phenomenon of *multiple* equilibria can coincide with a *qualitative* indeterminacy of concepts of optimal solution. As *SG* emphasizes quite rightly, the infinite regression of optimization belongs to another group of difficulties; for a first attempt at elucidation, cf. Mongin and Walliser (1987).

15. The confusion between "doing" and "bringing out" is perhaps related to that tradition in the philosophy of action that Ryle (1949) deprecated as the "intellectualist

myth" chap. 1): the thesis that intentionally directed action presupposes a *representation* of what has to be accomplished.

16. The ends/means model assumes that the agent has a more or less clear representation of how the action unfolds. That is not the case with the preferences/feasible set model which thus appears at once more general (from this standpoint) and, at the same time conducive to the confusion of "doing" and "bringing about." Harsanyi (1976, 92) argued for the second model's superior generality following a traditional line of argument (it would permit precise trade-offs between the obtained result and the expended means) which is debatable; for it assumes continuity of preferences, and this postulate was claimed to be unrelated to the raw notion of rationality.

17. The connection between "rationality" and "rational expectations" is explored in Walliser (1985) and Mongin (1989).

18. Perhaps controversially, Elster concludes that causal preference adaptation, unlike intentional adaption, most often implies exaggerated adjustments of the desirable to the possible. As explained in *SG*, the idea of *overshooting* comes from Veyne (1976), where it does not however serve to test the distinction between the causal and the intentional.

19. A *simultaneous* dependence on the feasible set and on the realized state is visible in the following reformulation of the Paris-London example: when the agent is in Paris and can no longer leave, the agent regrets London. The same complication underlies the 19th-century French saying: "Ah, que la République était belle du temps de l'Empire."

20. Elster speaks ambiguously of the "possible state z" (p. 134); but pages 135-36 suggest that it is a merely logical, and not material, possibility.

21. Note the new discussion of time preference in the second edition of *US*.

22. The objective control of which Weber was thinking takes the form of a *test* when the ideal-type (the external norm of interpretation, in history of philosophy) is confronted with actual behavior (resp. the text).

23. We have tried to counter the impossibility thesis in the earlier French version of this essay (sec. 6).

24. Reproductive aptitude is the maximandum that is normally attributed to natural selection, but *the human subject has no natural maximandum* if only because he can maximize at variable logical levels, a point which connects with the "infinite regression" structure inherent in any theory of rational choice.

25. Note that the use of this concept is important for morally appraising desires, but it is not sufficient: One might "want to distinguish heteronomous desires from unethical ones" (*SG*, 20). Elster thus distances himself once again from Kantianism.

26. Useful information arises, more generally, from the gap between reality and the norm: the case where the "behavior" and the "rational model" occupy these two abstract roles is after all but a privileged example. The Austrian theory of the business cycle sets the schema in action on the states of disequilibrium and equilibrium respectively.

27. Neither the Weberians nor the Popperians have come to grips with the duality that is suggested here of the intuitive notion of a "constraining model."

28. In parallel with the distinction between "sociological" and "economic" approaches, Elster suggests an original bifurcation: "There seems to be a choice between postulating partial or incomplete preference orderings, and postulating complete pref-

erences subject to endogenous change" (*SG*, 8). It may be feared, however, that the two types of models thus distinguished will account for the same givens equally well.

29. They would relate to the confusion, which is structural in the ideal-typical method, between the context of discovery and the context of refutation.

REFERENCES

Agassi, J. [1960] 1973. Methodological individualism. *British Journal of Sociology* 2: 244-70. Reprinted in *Modes of individualism and collectivism*, edited by J. O'Neill. London: Heinemann.

Aron, R. [1936] 1969. *La philosophie critique de l'histoire: Essai sur une théorie allemande de l'histoire*. 2d ed. Paris: Le Seuil.

Becker, G. S. 1976. *The economic approach to human behavior*. Chicago: University of Chicago Press.

Davidson, D. 1980. *Essays on actions and events*. Oxford: Oxford University Press.

Descartes, R. 1963. *Oeuvres philosophiques*, tome 1, edited by F. Alquié, Paris: Garnier.

Elster, J. 1975. *Leibniz et la formation de l'esprit capitaliste*. Paris: Aubier.

———. 1978. *Logic and society*. New York: Wiley.

———. [1979] 1986. *Ulysses and the sirens*. 2d ed. Cambridge: Cambridge University Press.

———. 1983a. *Explaining technical change*. Cambridge: Cambridge University Press.

———. 1983b. *Sour grapes*. Cambridge: Cambridge University Press.

———. 1985. *Making sense of Marx*. Cambridge: Cambridge University Press.

———. 1989a. *The cement of society*. Cambridge: Cambridge University Press.

———. 1989b. *Solomonic judgments*. Cambridge: Cambridge University Press.

Harsanyi, J. C. 1976. *Essays on ethics, social behavior and scientific explanation*. Dordrecht: Reidel, Theory and Decision Library.

———. 1977. *Rational behavior and bargaining equilibrium in games and social situations*. Cambridge: Cambridge University Press.

Hempel, C. G. 1965. *Aspects of scientific explanation*. New York: Free Press.

Hutchison, T. W. [1938] 1960. *The significance and basic postulates of economic theory*. Reprint. New York: Kelley.

Laqueur, W. 1980. *The terrible secret*. Boston: Little, Brown.

Merton, R. K. 1957. *Social theory and social structure*. 2d ed. Glencoe, IL: Free Press.

Mongin, P. 1986. Simon, Stigler et les théories de la rationalité limitée. *Information sur les sciences sociales/Social Science Information* 25:555-606.

———. 1988. La théorie du choix rationnel comme morale et psychologie. In *L'âge de la science: Ethique et philosophie politique*, 157-97. Paris: Odile Jacob.

———. 1989. Les anticipations rationnelles et la rationalité: Examen de quelques modèles d'apprentissage. Université Catholique de Louvain, Département des sciences économiques, Working Paper no. 8919.

Mongin, P., and B. Walliser. 1987. Infinite regression in the optimizing theory of decision. In *Risk, decision and rationality*, edited by B. Munier. Dordrecht: D. Reidel.

Popper, K. R. [1945] 1900. *The open society and its enemies*. 5th ed., Vol. 2. London: Routledge & Kegan Paul.

———. 1967. La rationalité et le statut du principe de rationalité. In *Les fondements philosophiques des systèmes économiques*, edited by E. M. Claassen, 142-150. Paris: Payot.

———. 1972. Die Logik der Sozialwissenschaften. In *Der Positivismusstreit in der deutschen Soziologie*, edited by T. W. Adorno. Darmstadt: Luchterhand.

Rosenberg, A. 1976. *Microeconomic laws*. Pittsburgh: University of Pittsburgh Press.

Ryle, G. 1949. *The concept of mind*. London: Hutchison.

Sen, A. K. 1970. *Collective choice and social welfare*. Amsterdam: North Holland.

Simon, H. A. 1983. *Models of bounded rationality*. 2 vols. Cambridge: MIT Press.

Tversky, A., and D. Kahneman. 1981. The framing of decisions and the psychology of choice. *Science* 211:453-58.

Veyne, P. 1976. *Le pain et le cirque*. Paris: Le Seuil.

Walliser, B. 1985. *Anticipations, équilibres et rationalité économique*. Paris: Calmann-Lévy.

Watkins, J.W.N. 1970. Imperfect rationality. In *Explanation in the behavioral sciences*, edited by R. Borger and F. Cioffi. Cambridge: Cambridge University Press.

Weber, M. 1922. *Gesammelte Aufsätze zur Wissenchafslehre*. Tübingen: Mohr. Partial English translation as *The methodology of the social sciences*. Glencoe, IL: Free Press, 1969.

Philippe Mongin is a senior research fellow with the Centre National de la Recherche Scientifique and the Ecole normale supérieure, Paris. He has visited the universities of Cambridge (UK), Montreal, Louvain-la-Neuve and Chicago. His work is concerned with the philosophy of the social sciences, the history of economics, and formal choice theory. He has published articles in a variety of French- and English-language journals and is currently working on a book, How Economists Explain.